An Entropy Based Approach to Real-Time Information Extraction for Industry 4.0

Marcello Trovati, Huaizhong Zhang, Jeffrey Ray and Xiaolong Xu

***Abstract*—Industry 4.0 has drawn considerable attention from industry and academic research communities.**

The recent advances in Internet of Things (IoT), Big Data Analytics, sensor technology and Artificial Intelligence have led to the design and implementation of novel approaches to take full advantage of data-driven solutions applicable to Industry 4.0. With the availability of large datasets, it has become crucially important to identify the appropriate amount of relevant information, which would optimise the overall analysis of the corresponding systems. In this article, specific properties of dynamically evolving data systems are introduced and investigated, which provide framework to assess the appropriate amount of representative information.

***Index Terms*—Industry 4.0, Entropy, Network Theory, Dynamical Systems**

I. Introduction

THE concept of Industry 4.0, or fourth industrial revolution, was recently introduced to describe the increasing need and motivation to achieve higher automation and data exchange in manufacturing technologies. These include cyber-physical systems, the Internet of Things, Cloud computing, Big Data Analytics and Artificial Intelligence [1]. In particular, Industry 4.0 is characterised by the full automation and digitisation processes in the manufacturing industry via the development of novel technologies, which will have deep and far reaching implications, especially in small and medium-sized enterprises [4]. The availability of large quantities of real-time data has created enormous opportunities, as well as critical challenges [5]. Therefore, the integration of novel big data analytics techniques with Industry 4.0 has been the focus of considerable research to fully utilise the available data [6].

The motivation of this work is to provide a novel framework to assess the level of *information saturation* with respect to a real-time data system relevant to Industry 4.0. Loosely speaking, this is defined as the least amount of real-time information, which is sufficient to fully describe the main properties related to the associated system. This is particularly useful in providing an efficient approach, which allows the evaluation of the level of information that is deemed sufficient. Furthermore, this facilitates an optimised information extraction process in a real-time context.

M. Trovati, H. Zhang and J. Ray are with the Department of Computer Science, Edge Hill University, UK; trovatim@edgehill.ac.uk.

X. Xu is with the School of Computer Science, Nanjing University of Posts and Telecommunications Nanjing, China.

The article is structured as follows: in Sections II and III the main concepts are introduced and discussed in the context of the existing methods and approaches. In Section IV, the main results are introduced, which are supported by the evaluation in Section VI. Finally, Section VII concludes the article.

II. Related Work

Industry 4.0 was motivated by the German manufacturing strategy focusing on maintaining and promoting its relevance in the industrial sector, by integrating technologies, new data-driven business models and their digital transformation [7]. Large volumes of data are continuously being generated from different information sources, whose analysis requires considerable resources [13], [14]. Advanced Data Analytics can address this challenge by facilitating the extraction of knowledge from huge datasets to enable manufacturers to carry out the evaluation of their products' lifecycle during the various stages [14]. In fact, the information shared by systems within an Industry 4.0 context creates connected and collaborative environments, which require prediction tools, so that data can be efficiently processed in near real-time to address uncertainties and identify actionable information and facilitate the decision making process [19].

A. Information Theory

Information theory consists of a variety of pure and applied disciplines including mathematical sciences, artificial intelligence, and complexity science, which is concerned with the use, transmission and assessment of the stochastic properties related to information. Historically, this was introduced by Shannon as the quality of information shared via a set of messages, which are affected by overall noise [15].

Entropy plays an essential role in information theory, whose goal is to assess the level of relevant and accurate information that is shared across one or more systems [2], [3]. Loosely speaking, this can be regarded as the level of useful information, which is actually shared. More specifically, entropy is the expected number of bits of shared information, over all its different combinations.

B. Network Theory

Networks have been used to model several complex systems, with applications to numerous research areas, such as mathematics, biology, psychology and sociology.
They are defined by a *node set* $V = \{v_i\}_{i=1}^{n}$, which contains

nodes that are connected by edges $e_{v_i,v_j} \in E$, denoted as the *edge set* [16]. We denote $\#V$ and $\#E$ the number of nodes and edges, respectively. In this article self-loops are not allowed, or in other words, $e_{v_i,v_i} \notin E$.
Networks have been demonstrated to provide modelling tools to address a variety of highly complex systems as they often consist of numerous simple entities, which are mutually connected [16].

The topology of different networks has been extensively investigated to identify crucial information on the corresponding system, which can provide a set of predictive tools to investigate its properties. In particular, stochastic topological features can successfully model systems based on unknown parameters. The most well-known types of such networks include, random and scale-free networks [20].

Random networks are usually described by a probability distribution p, which specifies the existence of the edges between any two nodes. Such probability is linked to the fraction p_k of nodes with degree k as follows:

$$p_k \approx \frac{z^k e^{-z}}{k!} \tag{1}$$

where $z = (n-1)p$, and n is the number of nodes.

A type of networks, which appears in numerous contexts, is scale-free networks [16]. These are characterised by the property that their node degrees are governed by a power law defined as

$$p_k \approx k^{-\gamma} \tag{2}$$

for large values of k, where γ has been demonstrated to be generally in the range $2 < \gamma < 3$, and p_k is the fraction of nodes with degree k. Other types of scale-free networks are those exhibiting a preferential attachment, which play an important role in network theory. This is a stochastic process, in which additional connections are added continuously to nodes in a directly proportional fashion with respect to the number of connections of the corresponding nodes [16]. The Barabási-Albert model is an algorithm based on random scale-free networks defined by a preferential attachment process. This type of network can be seen in various contexts such as the Internet, the World Wide Web, citation networks, and specific social networks [16].
This type of network is initially defined as a connected network with m_0 nodes and each new node is connected to $m \leq m_0$ existing nodes with a probability p_i to node i defined as

$$p_i = \frac{k_i}{\sum_j k_j}. \tag{3}$$

C. Text Mining within Industry 4.0

Text mining is a well-established set of tools, techniques and approaches aiming to extract relevant information from (unstructured) textual data sources. Text mining focuses on the accurate extraction of information, and its semantic properties, from unstructured textual data sources. However, the analysis and processing of information from text corpora is typically a complex task, depending on their sizes, sources and the level of ambiguity. There are a variety of techniques to achieve this, which include grammar-based text extraction defined by text patterns identifying text fragments with a specific structure [24].
The concept of Industry 4.0 focuses on methodologies, which are based on a multitude of data-driven approaches, where unstructured datasets contain crucial information [22]. Examples of text mining usage within Industry 4.0, include systematic reviews, business intelligence assessment and social media analysis [23].

III. Main Definitions

In this section, the main notions crucial to the concept of entropy and information saturation are discussed.

The main assumptions in this article are as follows:

1) Information is added in terms of new connections. In other words, the dynamics associated with a system is defined by adding new edges between nodes at each time iteration;
2) The number of nodes is constant. This is not unreasonable as there are many systems where this occurs. For example, totally disconnected nodes would be associated with data with no relevant information;
3) The connections are represented by weightless edges. This might be regarded as an over simplistic assumption. However, there are many real-time systems where weightless networks provide suitable predictive models [21]. Future investigation will focus on more general weighted networks.

More specifically, let $G_t = G_t(V_t, E_t)$ be a dynamic network with respect to time t, where $V_t = \{v_i\}_{i=1}^{\#V_t}$, and $E_t = \{e_{ij}\}_{i\neq j=1}^{\#V_t}$ are the node and edge sets, respectively. Define $n_t(v_i)$ as the number of neighbours of the node v_i at time t (or the degree of v_i), and $m_t = \#V_t - 1$. Note that is the number of nodes is constant (that is no new nodes are either added or removed) then m_t and V_t will be simply referred to as m and V, respectively.
The rate of change of connectivity of a node v_i is calculated as

$$\frac{n_{t+1}(v_i) - n_t(v_i)}{n_t(v_i)}. \tag{4}$$

We then define

$$\Delta_t(v_i) = \frac{|n_{t+1}(v_i) - n_t(v_i)|}{|m - n_t(v_i)|}, \tag{5}$$

which is the renormalised version of Equation 4. Note that if $n_{t+1}(v_i) = n_t(v_i)$ (no change in the connectivity of the node v_i), then $\Delta_t(v_i) = 0$ and if $n_{t+1}(v_i) = m$ (the node v_i is connected to every other node), then $\Delta_t(v_i) = 1$.

We define the entropy of the node v_i at time t as

$$H_t(v_i) = -\Delta_t(v_i) \log_{\#V} (\Delta_t(v_i)), \tag{6}$$

where $m = \#V - 1$, which is assumed to be constant over time. The motivation to use such logarithmic base is that the

number of nodes are regarded as the unit of information of the system introduced in this article.

We, therefore define the entropy of the network G_t at time t as

$$H_t(G_t) = -\sum_{i=1}^{\#V} \Delta_t(v_i) \log_{\#V}(\Delta_t(v_i)). \tag{7}$$

Finally, the entropy of the network G_t for $t = t_1, \ldots, t_l$ is defined as

$$\mathbf{H}_t(G_t) = (H_{t_1}(G_{t_1}), \ldots, H_{t_l}(G_{t_l})) \tag{8}$$

From Equation 7, we can see that $H_t(G_t) = 0$ if the network is complete or it has no change in its topology. In other words, small values of the entropy imply a small number of new edges added, or the network being very close to a complete topology. This is equivalent to a very limited amount of new information related to the network configuration. Note that for $v \in V$,

$$0 \leq H_t(v) \leq \frac{1}{e \ln(\#V)},$$

which implies that

$$0 \leq H_t(G_t) \leq \frac{\#V}{e \ln(\#V)}.$$

Definition 1: For a network $G_t = G_t(V, E_t)$ with entropy $\mathbf{H}_t(G_t)$, we say that G_t has reached *information saturation* at time T if $H_{t_k}(G_{t_k}) < \epsilon$ for all $k > T$ and $\epsilon > 0$.

A. An Integer Based Network Coding

In information theory, a *code*, which is defined via a set of *words*, can provide a precise yet efficient way to describe the states of the system.

In this section, we will introduce an integer based network coding, which uniquely identifies networks up to isomorphisms. The main benefit of this type of code is its ability to provide information based on specific properties of the dynamics of the corresponding system, as discussed in Section V.

Let $f : V \to \mathcal{P}_G$ be a bijective map, where $n_t(v_i) \geq n_{t+1}(v_i) \Rightarrow f(n_t(v_i)) \geq f(n_{t+1}(v_i))$ and let $\mathcal{P}_G = \{p_i\}_{i=1}^{\#V}$ be the set of increasing prime numbers or in other words, $p_i < p_{i+1}$. For example, for a network with 5 nodes, we have that $\mathcal{P}_G = \{2, 3, 5, 7, 11\}$. Recall that for a node v_i, $n(v_i)$ is the number of nodes adjacent to it. Define the ordered sequence

$$\mathcal{N}(G) = (n(v_j))_{j=1}^{\#V}, \tag{9}$$

such that $n(v_j) \leq n(v_k)$, for $j < k$ and $v_i, v_j \in V_t$

Definition 2: Let $G_t = G_t(V_t, E_t)$ be a network and let $\mathcal{N}(G)$ be as defined in Equation 9. We define the *integer code* of G_t as

$$\alpha(G_t) = \prod_{i=1}^{\#V} p_i^{n(v_i)} \tag{10}$$

For example, for the network depicted in Figure 1. Note that $\mathcal{N}(G) = \{1, 2, 2, 3, 3\}$. Therefore

$$\alpha(G) = 2^1 + 3^2 + 5^2 + 7^3 + 11^3 = 1710 \tag{11}$$

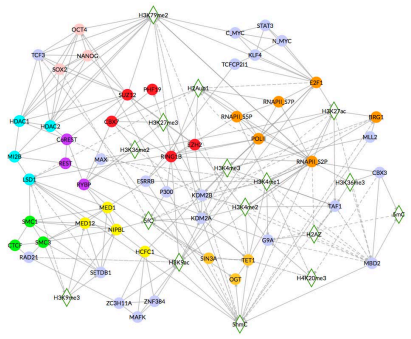


Fig. 1. Example of a complex network based on the epigenomic communication system of t5 embryonic stem cells [18].

From Definition 2, we have the following trivial result:

Lemma 1: Let $G_1 = G_1(V_{G_1}, E_{G_1})$ and $G_2 = G_2(V_{G_2}, E_{G_2})$. Then

$$\alpha(G_1) = \alpha(G_1) \Leftrightarrow G_1 \cong G_2 \tag{12}$$

Theorem 1:
Let $H_t(G_t)$ as defined in Equation 7. Then

$$H_t(G_t) \geq 2 \log_2 \left(\frac{\alpha(G_t)}{\alpha(G_{t+1})} \right) - \frac{1}{\#V^2} \tag{13}$$

Proof: We can see that

$$\begin{aligned} H_t(G_t) \quad &\geq \\ &- \sum_{i=1}^{\#V} \left(\Delta_t(v_i) + \frac{1}{\#V} \right)^2 = \\ &= -\log_2 2^{\sum_{i=1}^{\#V} \Delta_t(v_i)^2} \\ &- \log_2 2^{\frac{2}{\#V} \sum_{i=1}^{\#V} \Delta_t(v_i)} - \frac{1}{\#V^2} \\ &\geq -2 \log_2 2^{\sum_{i=1}^{\#V} \Delta_t(v_i)} - \frac{1}{\#V^2} \\ &\geq -2 \log_2 \frac{\alpha(G_{t+1})}{\alpha(G_t)} - \frac{1}{\#V^2} \\ &= 2 \log_2 \frac{\alpha(G_t)}{\alpha(G_{t+1})} - \frac{1}{\#V^2} \end{aligned} \tag{14}$$

■

Corollary 1: If $H_t(G_t) < \epsilon$, then

$$\alpha(G_{t+1}) > \alpha(G_t) 2^{-\frac{\epsilon \#V^2 + 1}{2\#V^2}} \tag{15}$$

Proof: This follows from Theorem 1. ■

IV. Main Results

In this section the main results based on the concepts and definitions introduced above are discussed, which will be differentiated between *structure-less* networks (purely random networks, whose edges are added based on an unknown probability P) and networks which follow a preferential attachment defined by the Albert - Barabàsi model.

A. Structure-less Networks

When no information is available on a networks, its topology or dynamical behaviour, any prediction of its properties is likely to be less accurate and specific.
Let

$$\Lambda_{\epsilon,\beta} = \{v \in V : H_{T-j}(v) < \epsilon, \text{ for } \epsilon > 0, j = 0, \dots, \beta\} \tag{16}$$

In other words, $\Lambda_{\epsilon,\beta}$ is the set of nodes such that $H_T(v), \dots, H_{T-\beta}(v)$ are close enough to 0, with respect to some $\epsilon > 0$. However, Equation 16 is not sufficient to guarantee information saturation, as the entropy should also exhibit a decreasing trend.
Since the edges are added randomly (as opposed to, for example, preferential attachment), G_t is information saturated with probability

$$P_{\epsilon,\beta}(G_t) = \frac{\#\Lambda_{\epsilon,\beta}}{m_t}. \tag{17}$$

As Equation 16 must follow a decreasing trend, due to the limited amount of specific information on the creation of new edges, linear regression is applied to the sequence $(P_{\epsilon,\beta}(G_t))_{t=1}^{k}$ to estimate its behaviour. More specifically, this will generate a linear regression line of the form $P_{\epsilon,\beta}(G_t) = a + bt$, for $a, b \in \mathbb{R}$. This is by no means the only approach, which could be used in this context, as there are several ways to assess the trend of the corresponding data. However, in this article it was noted that linear regression suffices to identify the overall data trend.

B. Preferential Attachment Based on Albert - Barabási Model

As discussed in Section II-B, the majority of real-world networks tend to exhibit specific behaviours and dynamical patterns, which have been extensively investigated [20]. In particular, there are numerous instances of preferential attachment properties, which have been observed both empirically and theoretically in many stochastic networks. In other words, new edges or nodes are likely to be added to highly connected nodes.
As discussed above, we assume new information is incrementally added in term of new connections (edges) between nodes. However, we might not know how and more importantly, where these new connections are added. As a consequence, we can only base our prediction on specific topological and dynamical properties of the corresponding networks. In this article, we will consider the Albert - Barabási preferential attachment model as the main underlying property characterising them, and in particular, this is defined as

$$P(v) = \frac{n_t(v)}{\sum_{v \in V} n_t(v)}, \tag{18}$$

where $n_t(v)$ is the degree of the node v at time t, and $p(v)$ is the probability of v having a new edge at time $t+1$.

Consider a node $v \in V$. Its entropy is minimised whenever

1) $\Delta_t(v) \approx 0$, or it is very small;
2) $\Delta_t(v) \approx 1$

Theorem 2: Let $\epsilon > 0$ and

$$\epsilon = \frac{\tilde{\epsilon}}{\#V}.$$

Then

$$P(H_t(v) < \epsilon) = 1 - \sum_{i=1}^{\lceil (m-n_t(v))(1-\epsilon) \rceil} i \left(\frac{n_t(v)}{\sum_{v \in V} n_t(v)} \right) \tag{19}$$

and

$$P(H_t(G) < \tilde{\epsilon}) = P\left(\sum_{v \in V} H_t(v) < \tilde{\epsilon} \right).$$

Proof: Note that

$$H_t(v) = -\Delta_t(v) \log\left(\Delta_t(v)\right) < 1 - \Delta_t(v) \tag{20}$$

based on its Taylor's polynomial expansion. Define $\epsilon > 0$ and assume that $H_t(v) < \epsilon$. If

$$1 - \Delta_t(v) < \epsilon \Rightarrow \Delta_t(v) > 1 - \epsilon$$

then this will guarantee $H_t(v) < \epsilon$.
Recall that

$$\Delta_t(v) = \frac{n_{t+1}(v) - n_t(v)}{m - n_t(v)}.$$

Thus,

$$\begin{aligned} &\frac{n_{t+1}(v) - n_t(v)}{m - n_t(v)} > \\ &1 - \epsilon \Rightarrow n_{t+1}(v) - n_t(v) > \\ &\qquad (m - n_t(v))(1 - \epsilon). \end{aligned} \tag{21}$$

This implies that

$$\begin{aligned} P(H_t(v) \;&<\; \epsilon) = P(n_{t+1}(v) - n_t(v) > (m - n_t(v))(1-\epsilon) \\ &= 1 - P(n_{t+1}(v) - n_t(v) < \lceil (m - n_t(v))(1-\epsilon) \rceil) \\ &= 1 - \sum_{i=1}^{\lceil (m-n_t(v))(1-\epsilon) \rceil} i \left(\frac{n_t(v)}{\sum_{v \in V} n_t(v)} \right) \end{aligned} \tag{22}$$

The equations above only consider a single node v. In particular, assume we want to evaluate the probability of $H_t(G) < \tilde{\epsilon}$

$$P(H_t(G) < \tilde{\epsilon}) = P\left(\sum_{v \in V} H_t(v) < \tilde{\epsilon} \right) \tag{23}$$

Assume

$$\epsilon = \frac{\tilde{\epsilon}}{\#V},$$

which will satisfy Equation 23. ∎

1) Entropy Assessment Based on Integer Coding: We will consider the a similar approach based on the integer coding described in Section III-A

Lemma 2: Based on the notation described in Theorem 2, we have that if $P(H_t(v) < \epsilon)$, then

$$\begin{aligned} P\left(\frac{\alpha(G_{t+1})}{\alpha(G_t)} \geq p_{k_{\lceil (m-n_t(v))(1-\epsilon) \rceil}} \right) = \\ \frac{n_t(f^{-1}(p_{k_{\lceil (m-n_t(v))(1-\epsilon) \rceil}}))}{\sum_{v \in V} n_t(v)} \end{aligned} \tag{24}$$

Proof: This result follows from Theorem 2

∎

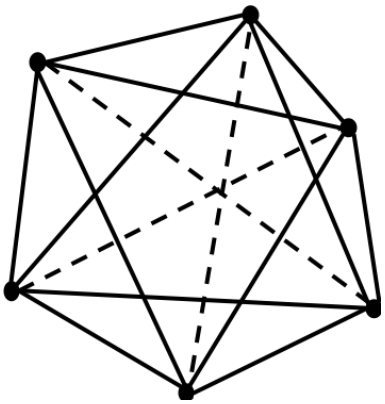


Fig. 2. Example of a $3-$missing network with 6 nodes.

V. Entropy as a Dynamical System

From the above discussion, we can easily see that the sequence

$$G_{t_1}, G_{t_2}, \ldots, G_{t_T} \tag{25}$$

defines a dynamical system on the topology of the network (recall that the node set is not changed and only edges are affected). More specifically, it is also trivial to see that if we consider the scenario that $\#E(G_{t_i}) \leq \#E(G_{t_j})$ for $i < j$, then

$$\alpha(G_{t_{i+1}}) = \beta_i\, \alpha(G_{t_i}), \tag{26}$$

where

$$\beta_i = \prod_{k \in K} p_k, \quad \text{for some } K \subset \mathbb{N} \text{ and } p_k \in \mathcal{P}_G. \tag{27}$$

As a consequence, we can identify a dynamical system equivalent to (25) as follows

$$(\beta_i)_{i=1}^{T}, \tag{28}$$

which captures the topology dynamics at each time iteration. In order to investigate some of the properties of its dynamics, we need to introduce the following definition

Definition 3: A network G_t is said to be $n-$complete if it has n missing edges to be complete.

See Figure 2 for an example of a $3-$missing network with 6 nodes.

Recall that G_t is complete if and only if $\alpha(G) = \prod_{k=1}^{\#V} p_k^{\#V-1}$.

Let

$$\mathcal{D}(\beta_i) = \frac{1}{2} \sum_{p_k} k, \tag{29}$$

that is the sum of all the exponents of all the prime numbers as in Equation 27. We then have the following trivial result

Lemma 3: A network G is $n-$complete if $\mathcal{D}(\beta_i) = n$.

From Equation 29, we can also see that the dynamical system defined in (28) is equivalent to

$$(\mathcal{D}(\beta_i)))_{i=1}^{T}. \tag{30}$$

Note that if $\#E(G_{t_i}) \leq \#E(G_{t_j})$ for $i < j$, then

$$\beta_{t_i} \geq \beta_{t_j} \Rightarrow \mathcal{D}(\beta_{t_i}) \geq \mathcal{D}(\beta_{t_j}),$$

as the network G_t asymptotically tend to completeness.

Definition 3 does not specify which edges are missing from the network in order to be complete. The most trivial case is they are all incident to the same node v, which implies that $\deg(v) = (\#V - 1) - n = m - n$. However, this is not true in general as they might be spread across the network. In other words, there is a set $\{\gamma_i\}_{i=1}^{h}$ (the missing edges would be incident to h nodes) such that $\sum_{i=1}^{h} \gamma_i = 2n$, and

$$\beta_t = \prod_{i=1}^{h} p_i^{m-\gamma_i-n_t(v_i)}. \tag{31}$$

In particular, the corresponding entropy is then calculated as in Equation 6, where the different $\Delta_t(v_i)$ are as follows

$$\Delta_t(v_i) = \frac{m - \gamma_i - n_t(v_i)}{m - n_t(v_i)}. \tag{32}$$

Note that for an $n-$ missing network, the only case where there is only one missing edge from two or more nodes occurs when the number of all nodes $\#V$ is an even number and $n \leq \dfrac{\#V}{2}$.

Theorem 3: For a network based on the preferential attachment based on the Albert - Barabási model associated with the dynamical system defined in (30), if $H_t(G_t) < \epsilon$, we have that

$$\beta_{t,\epsilon} = \prod_{i=1}^{h} p_i^{(m-n_t)(2-\epsilon)}. \tag{33}$$

Proof: This is directly derived from Theorem 2. ■

VI. Experimental Evaluation

This section will discuss the assessment of the model introduced in the previous sections. The evaluation consists of three different stages. The first one is based on two synthetic datasets, which were created to mathematically simulate and evaluate the different concepts and results described above. The second stage focuses on the analysis of a large textual dataset, identifying concepts related to Industry 4.0. This was utilised to assess the level of information associated by mutual relationships, as discussed in Section VI-B. Finally, Section VI-C describes the analysis carried out on the dataset described in [25], which is a dynamic social network created from an online community for students at University of California, Irvine.

A. Mathematical Simulation

In this part of the evaluation, two networks were created as follows. The first, which is depicted in Figure 3, is a randomly generated network, and new edges were added based on a probability distribution. The second network, as in Figure 4, was created based on the Albert-Barabási preferential

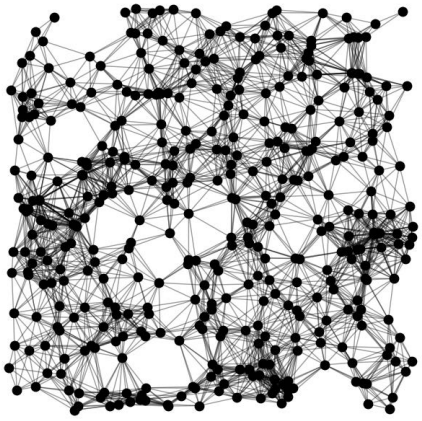


Fig. 3. The random network described in Section VI.

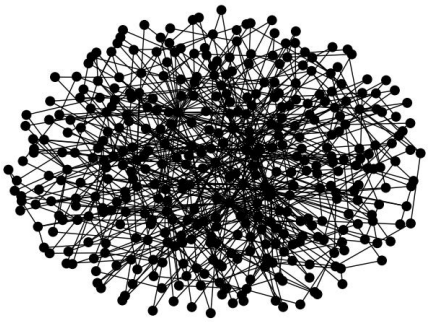


Fig. 4. The network based on the Albert-Barabási preferential attachment model as described in Section VI.

attachment model [16]. In more details, both networks had 200 nodes and the former had its dynamics defined via time iterations, during which a random number of edges (between 0 and 50 edges) were iteratively added, based on the preferential attachment law described by Equation 18. Subsequently, their entropy was evaluated as depicted in Figures 5 and 6.

Based on the method discussed in Section IV-B, the behaviour of the structure-less and the Albert-Barabási preferential attachment networks (as in Figures 3 and 4) was investigated for $\epsilon = 0.8$, and $\epsilon = 0.4$. The corresponding linear regression lines were calculated to be $P_{\epsilon,\beta}(G_t) = 0.48 - 0.0004t$ and $P_{\epsilon,\beta}(G_t) = 0.24 - 0.0003t$, respectively, as depicted in Figures 7 and 8.

This demonstrates that the dynamical behaviour of these networks, despite being based on a mathematical simulation, is consistent with the theoretical framework introduced above.

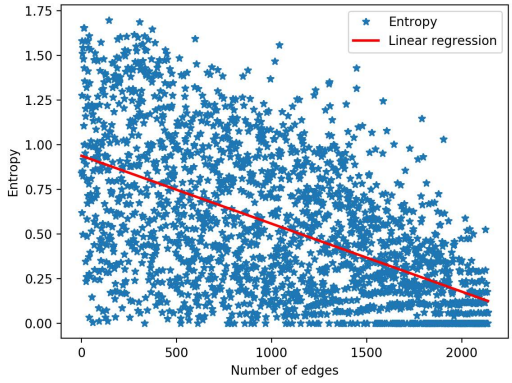


Fig. 5. The entropy based on the mathematical simulation of a structure-less dynamical network, described in Section VI. The red line indicates the linear regression, which shows a downward trend. This is to be expected as the more edges are added, the more the network will be close to being complete, which implies a decrease in entropy.

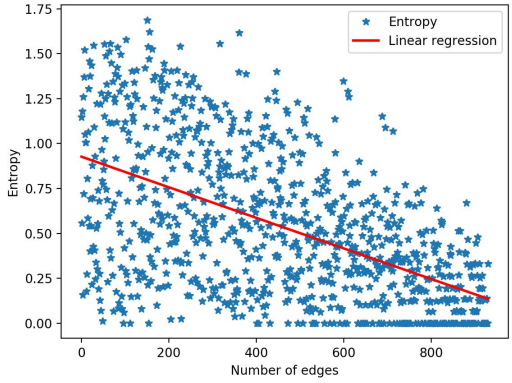


Fig. 6. The entropy based on the mathematical simulation of a dynamical network defined by the Albert-Barabási preferential attachment model, as described in Section VI. The red line indicates the linear regression, as described in Figure 5.

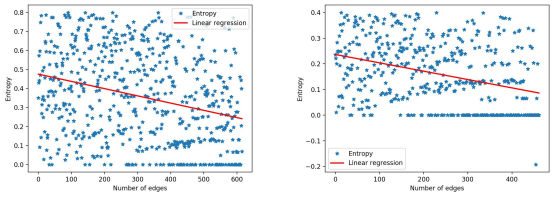


Fig. 7. The behaviour of the system associated with the structure-less network as in Figure 3, with $\epsilon = 0.8$, and $\epsilon = 0.4$, whose linear regression lines are defined as $P_{\epsilon,\beta}(G_t) = 0.48 - 0.0004t$ and $P_{\epsilon,\beta}(G_t) = 0.24 - 0.0003t$, respectively

B. Evaluation via Text Mining Information Extraction

The second part of the evaluation was based on the automated extraction of concepts related to Industry 4.0. In particular, a grammar-based approach was utilised and specific text fragments were identified that capture a relationship between relevant concepts [24]. The concepts and their mutual relations identified via text patterns, need to appear within the

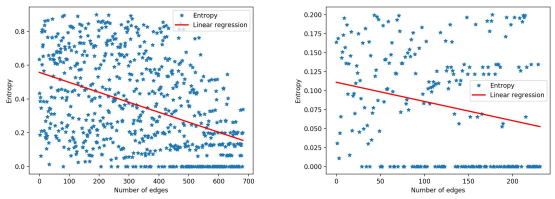


Fig. 8. The behaviour of the Albert-Barabási preferential attachment network as in Figure 4, with an $\epsilon = 0.8$, and $\epsilon = 0.4$ and linear regression line defined by $P_{\epsilon,\beta}(G_t) = 0.56 - 0.0006t$ and $P_{\epsilon,\beta}(G_t) = 0.11 - 0.0003t$, respectively

same text fragments, such as paragraphs. This was carried out by considering the following quintuples `(NP1, Rel, NP2)` where:

- `NP1` and `NP2` are the *noun phrases*, i.e. phrases with a noun as the head word.
- `Rel` refers to verbs, cue phrases and keywords related to specific dependency and causal relationships linking the concepts.

As discussed in [21], a grammar-based approach has been demonstrated to be one of the mot efficient and accurate ways to successfully extract relationships between concepts.
Subsequently, these triples were extracted by investigating the syntactic structure of sentences and text fragments. Python NLTK [26] was used to tokenise, parse and extract the relevant syntactic information which are described by the different POS components [24].
Approximately 400 freely available articles and specialised blogs based on Industry 4.0 were identified as follows:

- Suitable combinations of the following keywords
  - Industry 4.0
  - Automation
  - Manifacturing
  - Process efficiency

  were used to identify the above textual sources, which were downloaded in `xml` format.
- These were POS tagged, and analysed via text patterns as described above.
- The output of the above analysis was a list of two-ples

  `[concept_1, concept_2]`

  where `concept_1` and `concept_2` refer to concepts related to Industry 4.0

The triples `(NP1, Rel NP2)`, naturally create a network, where the concepts in `NP1` and `NP2` are connected by an edge whose type depends on the associated relations. In order to avoid duplication and unnecessary redundancy, similar (based on synonymity) concepts refer to the same node. Table I shows a small selection of such concepts. The resulting network had over 120 nodes and $3,000$ edges. To simulate the dynamics related to the network, 100 edges were sequentially added at each time iteration.

Figure 9 depicts the behaviour of the entropy of the dynamical network extracted via the Text Mining process, as described above.

TABLE I
SELECTION OF CONCEPTS IDENTIFIED BY THE TEXT MINING PROCESS DISCUSSED IN SECTION VI-B.

Concepts
Advanced Manufacturing
Data Science
Data-driven methods
Artificial Intelligence
German government
Mobile technologies
Internet of Things
Computer interfaces
Cyber security
Sensor Technology
Algorithms
Virtual reality

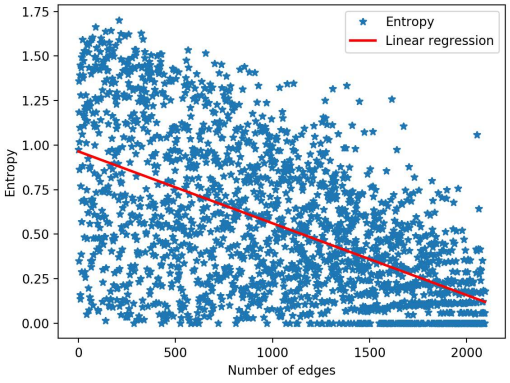


Fig. 9. The entropy values for the network extracted via the Text Mining process discussed in Section VI-B.

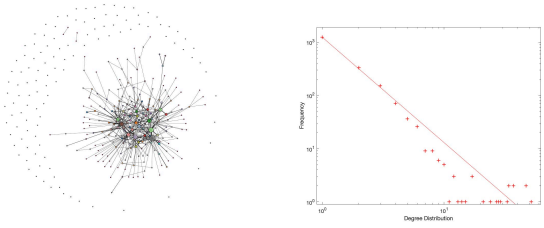


Fig. 10. The social network dataset introduced in [25].

C. Social Network Dataset

As described above, this dataset captures the user interactions across a social network, where nodes are individuals mutually connected by edges if at least one message is sent or received. This creates a dynamical network with approximately $1,900$ nodes and almost $60,000$ edges (stored chronologically), as depicted in Figure 10, which also shows the degree distribution, which is consistent with the Barabási - Albert model [16].

Subsequently, the approach described in Section IV-B was carried out to assess the overall entropy of this network, with respect to the corresponding time stamps. Figure 11 depicts

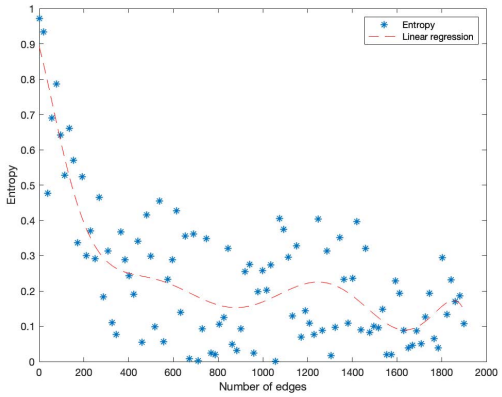


Fig. 11. The entropy values for the network described in Section VI-C, which includes a simple polynomial fit of the data values to emphasise the decreasing entropy values.

the numerical results, where a polynomial fit is included to demonstrate the decreasing nature of its behaviour.

VII. Conclusion

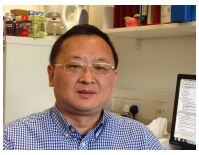
In this article, we have introduced a novel approach to assess the information saturation level of a real-time complex system, associated with a dynamical network. This allows the identification of the appropriate level of relevant information, which enables an efficient approach to information extraction and assessment. The evaluation was carried out based on the motivation to provide a valuable framework for Industry 4.0. Despite only weightless networks were considered, the experimental evaluation has demonstrated the potential of the proposed approach.

However, we acknowledge that some of the aspects introduced in this article need further research to fully take advantage of their theoretical properties. In particular, we are aiming to expand our method by including general weighted networks, as well as develop an algebraic topology analysis of the integer coding introduced in Section III-A. Preliminary results suggest that this can provide a fast and efficient framework in real-time data analysis particularly relevant to Industry 4.0.

References

[1] Hermann M.; Pentek T.; Otto B. 2016: Design Principles for Industrie 4.0 Scenarios, 2016 49th Hawaii International Conference on System Sciences (HICSS), 2016
[2] Pathria, R. K.; Beale, P. Statistical Mechanics (Third Edition). Academic Press. p. 51. ISBN 978-0123821881, 2011
[3] Thomas M. Cover; Joy A. Thomas, Elements of Information Theory. Hoboken, New Jersey: Wiley, 2006
[4] Sommer, L. Industrial revolution—Industry 4.0: Are German manufacturing SMEs the first victims of this revolution? Journal of Industrial Engineering and Management, 8, 1512-1532, 2015.
[5] R. Drath, A. Horch. Industrie 4.0: Hit or Hype?, IEEE Industrial Electronics Magazine, vol. 8, no. 2, pp. 56-58, 2014.
[6] H. Lasi, P. Fettke, H.-G. Kemper, T. Feld, M. Hoffmann. Industry 4.0, Business and Information Systems Engineering, vol. 6, no. 4, pp. 239-242, 2014.
[7] E. Hofmann, M. Rüsch. Industry 4.0 and the current status as well as future prospects on logistics, Computers in Industry 89 (2017) 23?34
[8] Alcácer, V. Cruz-Machado, Scanning the Industry 4.0: A Literature Review on Technologies for Manufacturing Systems, Engineering Science and Technology, an International Journal, Volume 22, Issue 3, 2019, Pages 899-919.
[9] F.-E. Bordeleau, E. Mosconi, L. A Santa-Eulalia, Business Intelligence in Industry 4.0: State of the art and research opportunities. Proceedings of the 51st Hawaii International Conference on System Sciences (2018)
[10] P. Marcon, et al. Communication Technology for Industry 4.0, 2017 Progress in Electromagnetics Research Symposium-Spring (PIERS) St. Petersburg, 2017 (2017), pp. 1694-1697,
[11] A. Rojko Industry 4.0 Concept: Background and Overview Int. J. Interactive Mobile Technol., 11 (2017), p. (5).
[12] F. Zezulka, P. Marcon, I. Vesely, O. Sajdl, Industry 4.0 - An Introduction in the phenomenon, IFAC-PapersOnLine 49 (25) (2016) 8?12.
[13] M. Bortolini, E. Ferrari, M. Gamberi, F. Pilati, M. Faccio, Assembly system design in the Industry 4.0 era: a general framework, IFAC-PapersOnLine 50 (1) (2017) 5700?5705.
[14] Q. Qi, F. Tao Digital Twin and Big Data Towards Smart Manufacturing and Industry 4.0: 360 Degree Comparison IEEE Access 6 (2018), pp. 3585-3593 10.1109/ACCESS.2018.2793265
[15] Shannon, C. E. A Mathematical Theory of Communication. Bell System Technical Journal. 27 (4): 623–666, 1948
[16] Albert R and Barab´asi A L. Statistical Mechanics of Complex Networks. Reviews of Modern Physics 74, 47, 2002
[17] Trovati M, Win T Y, Sun Q, and Kontonatsios G. Assessment of Security Threats via Network Topology Analysis: An Initial Investigation. Green, Pervasive, and Cloud Computing, 10232. pp. 416-425. ISSN 0302-9743, 2017
[18] David Juan, Juliane Perner, Enrique Carrillo de Santa Pau, Simone Marsili, David Ochoa, Ho-Ryun Chung, Martin Vingron, Daniel Rico, Alfonso Valencia. Epigenomic Co-localization and Co-evolution Reveal a Key Role for 5hmC as a Communication Hub in the Chromatin Network of ESCs. Cell Reports (2016).
[19] Jay Lee, Hung-An Kao, Shanhu Yang, Service Innovation and Smart Analytics for Industry 4.0 and Big Data Environment, Procedia CIRP, Volume 16, 2014, Pages 3-8, ISSN 2212-8271.
[20] Newman M. Networks: An Introduction, Oxford University Press, Inc. New York, NY, USA, 2010
[21] Trovati M, Hayes J, Palmieri F and Bessis N. Automated extraction of fragments of Bayesian networks from textual sources. Applied Soft Computing, 2017
[22] Oliff H, Liua Y. Towards Industry 4.0 Utilizing Data-Mining Techniques: A Case Study on Quality Improvement. The 50th CIRP Conference on Manufacturing Systems, 2017
[23] Yilmaz G., Aygün D. and Tanrikulu, Z. Social Media?s Perspective on Industry 4.0: A Twitter Analysis. Social Networking, 6, 251-261, 2017
[24] Manning C D and Schütze H. Foundations of Statistical Natural Language Processing. The MIT Press, 1999
[25] Rossi R, and Ahmed N. The Network Data Repository with Interactive Graph Analytics and Visualisation. http://networkrepository.com, AAAI, 2015.
[26] Bird S, Loper E, and Klein E. Natural Language Processing with Python. O'Reilly Media Inc., 2009

Marcello Trovati Marcello Trovati is a Reader in Computer Science in the Department of Computer Science, Edge Hill University. His research interests include mathematical modelling, data science, including data and text mining, and their applications to multi-disciplinary topics.



Huaizhong Zhan Huaizhong Zhang is a senior lecturer in the Department of Computer Science at Edge Hill University. His research interests focus on deep learning and big data, image analysis, pattern recognition, object tracking and recognition. In addition, he has extensive development experience, including Proteomics/open source pipelines, video surveillance system, and computer assisted instruction.

Jeffrey Ray Jeffrey Ray is a PhD student at the Department of Computer Science, Edge Hill University. He specialises in Big Data Analytics, Machine Learning, Natural language Processing and Artificial Intelligence

Xiaolong Xu Xiaolong Xu is currently a professor in the School of Computer Science, Nanjing University of Posts and Telecommunications, Nanjing 210003, China. He is also working for the Jiangsu Key Laboratory of Big Data Security and Intelligent Processing. He is a senior member of China Computer Federation. He teaches graduate courses and conducts research in areas of Cloud Computing, Big Data, Information Security and Novel Network Computing Technologies. As the leader of project teams, he has successfully completed a number of high-level research projects, including the projects sponsored by the National Science Fund of China. He has published more than 100 Journal and conference papers as the first or corresponding author and 5 books. He is authorized 52 patents by the State Intellectual Property Office of China as the first inventor. He was rated as excellent young professor of Jiangsu Province in 2014, selected as the high-level creative talents of Jiangsu province in 2015, and won the title of outstanding expert in the area of computer science and technology.